LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-JRNL-219959

A Kalman-filter bias correction of ozone deterministic, ensemble-averaged, and probabilistic forecasts

L. Delle Monache, G. A. Grell, S. McKeen, J. Wilczak, M. O. Pagowski, S. Peckham, R. Stull, J. McHenry, J. McQueen

March 21, 2006

Tellus

Updated October 14, 2003

A Kalman-filter bias correction of ozone deterministic, ensemble-averaged, and probabilistic forecasts

Luca Delle Monache[1,2], George Grell[3,4], Stuart McKeen[4,5], James Wilczak[6], Mariusz Pagowski[3,7], Steven Peckham[3,4], Roland Stull[1], John McHenry[8], Jeffery McQueen[9]

[1] *Atmospheric Science Programme, Earth and Ocean Sciences Department, University of British Columbia, Vancouver, British Columbia, Canada*

[2] *Now at Lawrence Livermore National Laboratory, Livermore, California, USA*

[3] *Global Systems Division, Earth System Research Laboratory, National Oceanic and Atmospheric Administration, Boulder, Colorado, USA*

[4] *Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado, USA*

[5] *Chemical Sciences Division, Earth System Research Laboratory, National Oceanic and Atmospheric Administration, Boulder, Colorado, USA*

[6] *Physical Sciences Division, Earth System Research Laboratory, National Oceanic and Atmospheric Administration, Boulder, Colorado, USA*

[7] *Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, Colorado, USA*

[8] *Baron Advanced Meteorological Systems, c/o North Carolina State University, Raleigh, North Carolina, USA*

[9] *National Weather Service / National Centers for Environmental Prediction / National Oceanic and Atmospheric Administration, Camp Springs, Maryland, USA*

Corresponding author address: Luca Delle Monache, Lawrence Livermore National Laboratory, 7000 East Avenue, L-103, Livermore, CA 94550, USA

E-mail address: ldm@llnl.gov

Abstract. Kalman filtering (KF) is used to postprocess numerical-model output to estimate systematic errors in surface ozone forecasts. It is implemented with a recursive algorithm that updates its estimate of future ozone-concentration bias by using past forecasts and observations. KF performance is tested for three types of ozone forecasts: deterministic, ensemble-averaged, and probabilistic forecasts. Eight photochemical models were run for 56 days during summer 2004 over northeastern USA and southern Canada as part of the International Consortium for Atmospheric Research on Transport and Transformation New England Air Quality (AQ) Study. The raw and KF-corrected predictions are compared with ozone measurements from the Aerometric Information Retrieval Now data set, which includes roughly 360 surface stations. The completeness of the data set allowed a thorough sensitivity test of key KF parameters. It is found that the KF improves forecasts of ozone-concentration magnitude and the ability to predict rare events, both for deterministic and ensemble-averaged forecasts. It also improves the ability to predict the daily maximum ozone concentration, and reduces the time lag between the forecast and observed maxima. For this case study, KF considerably improves the predictive skill of probabilistic forecasts of ozone concentration greater than thresholds of 10 to 50 ppbv, but it degrades it for thresholds of 70 to 90 ppbv. Moreover, KF considerably reduces probabilistic forecast bias. The significance of KF-postprocessing and ensemble-averaging is that they are both effective for real-time AQ forecasting. KF reduces systematic errors, whereas ensemble-averaging reduces random errors. When combined they produce the best overall forecast.

1. Introduction

The skill of ozone deterministic forecasts can be improved using ensemble methods [*Delle Monache and Stull*, 2003; *McKeen et al.*, 2005; *Delle Monache et al.*, 2005a], by combining weighted ensemble averaging with the application of linear regression [*Pagowski et al*., 2005a] or dynamic linear regression [*Pagowski et al.*, 2005b], and with bias removal methods [*McKeen et al.*, 2005; *Wilczak et al*., 2005; *Delle Monache et al*. 2005b].

Forecast bias, i.e., systematic error, is a problem common to all Chemistry Transport Models (CTMs) [*Russel and Dennis*, 2000]. This study evaluates the ability of the Kalman filter (KF) predictor post-processing bias-removal method in predicting biases of surface ozone forecasts. The KF correction is an automatic post-processing method that uses the recent past observations and forecasts to estimate the model bias in the future forecast, where bias is defined as the "difference of the central location of the forecasts and the observations" [*Jolliffe and Stephenson*, 2003].

The data set used in this study to test the KF has been collected during the International Consortium for Atmospheric Research on Transport and Transformation (ICARTT) New England Air Quality (AQ) Study. The experiment, including both ozone surface and upper air observations and predictions (archived by the National Oceanic and Atmospheric Administration (NOAA) Earth System Research Laboratory), was held during summer of 2004 over northeastern USA and southern Canada. The following eight CTMs (as described also in Table 1) were run from 0000 UTC 6 July to 0000 UTC 30 August 2004 (i.e., 56 days):

- A unified Regional Air-quality Modeling System (AURAMS, *Moran et al.* [1997]) and the Canadian Hemispheric and Regional Ozone and NO_x System (CHRONOS, *Pudykiewicz et al.* [1997]) provided by the Meteorological service of Canada.
- The Baron Advanced Meteorological System Multi-scale Air Quality Simulation Platform [*McHenry et al.*, 2005], run at 15 km (BAMS-15) and 45 km (BAMS-45), provided by Baron Advanced Meteorological System Inc. Corporation.
- The Community Multi-scale Air Quality Model (CMAQ/ETA, *Byun and Ching* [1999]) from the National Weather Service (NWS)/National Center for Environmental Prediction (NCEP).
- The Weather Research and Forecast Model/Chemistry model [*Grell et al.*, 2005] run with two different versions (version 1.3 (WRF/CHEM-1) and version 2.03 (WRF/CHEM-2)) by the NOAA Global Systems Division. WRF/CHEM is an on-line CTM, where the chemistry is fully coupled with the meteorology.
- The Sulfur Transport and Emissions Model (STEM, *Carmichael et al.* [2003]) provided by University of Iowa.

Hourly averaged surface ozone concentrations were available at roughly 360 stations and stored in the Aerometric Information Retrieval Now (AIRNow, http://www.epa.gov/airnow) database. The model domains, their overlap, and the station characterizations are shown in Figure 1. Further details about each model and the observation data can be found in *McKeen et al.* [2005].

Delle Monache et al. [2005b] showed that the KF-corrected forecasts are improved for correlation, gross error, root mean square error (RMSE), and unpaired peak prediction accuracy (UPPA). Their successful results prompted this extended study. The KF

method and algorithm are described in section 2. In section 3 a sensitivity analysis for one of the key filter parameters, the error ratio, is presented. An optimal value for this parameter is found by evaluating the KF performance in different situations with different meteorology and different AQ scenarios. With the error-ratio optimal value found, the filter performance is tested for deterministic, ensemble-averaged (section 4) and probabilistic surface ozone forecasts (section 5). In section 6 conclusions are drawn from those results.

2. Kalman Filter

The KF has been used in data-assimilation schemes to improve the accuracy of the initial conditions for numerical weather prediction (NWP) [e.g., *Burgers et al.*, 1998; *Hamill and Snyder*, 2000; *Houtekamer and Mitchell*, 2001; *Houtekamer et al*., 2005] and AQ forecasts [e.g., *van Loom et al.*, 2000; *Segers et al.*, 2005]. The KF has also been used for weather and AQ (i.e., ozone) forecasts as a predictor bias-correction method during post-processing of short-term forecasts [*Homleid*, 1995; *Roeger et al.*, 2003; *Delle Monache et al*., 2005b]. This latter approach is applied here. The filter uses a recursive algorithm to estimate the systematic component of the forecast errors, which often corrupts AQ forecasts [e.g., *Russel and Dennis*, 2000; *Delle Monache et al*., 2005b], thus effectively reducing bias.

The KF predictor-corrector approach is linear, adaptive, recursive and optimal. Namely, it predicts the future bias with a linear relationship, given by the old bias plus a quantity proportional to the difference between the verifying bias and the previous prediction. It differs from a neural-network approach, which is non-linear [e.g., *Cannon and Lord*, 2000]. While a neural-network approach requires a long training period and then statically produces a prediction, at each iteration the KF approach adapts its coefficients, resulting in a much shorter training period. However, KF is unable to predict a large bias when all biases for the past few days have been small. Finally, it is recursive because at any iteration values of the KF coefficients depend on the values at the previous iteration, and it is optimal in a least-square-error sense [*Delle Monache et al*., 2005b].

2.1 Filter Algorithm

Kalman [1960] wrote an algorithm based on the minimization of the expected mean-square error (p), computed a follows:

$$p_{t|t-\Delta t} = \left(p_{t-\Delta t|t-2\Delta t} + \sigma_\eta^2\right)\left(1 - \beta_{t|t-\Delta t}\right) \tag{1}$$

where $t \mid t-\Delta t$ means that the value of the variable at time t depends on values at time $t-\Delta t$, and β is a weighting factor, called the Kalman gain, which can be calculated from:

$$\beta_{t|t-\Delta t} = \frac{p_{t-\Delta t|t-2\Delta t} + \sigma_\eta^2}{\left(p_{t-\Delta t|t-2\Delta t} + \sigma_\eta^2 + \sigma_\varepsilon^2\right)} \tag{2}$$

The true (unknown) forecast bias x_t is modeled at time t, by the previous true bias plus a white noise η term [*Bozic*, 1994]:

$$x_{t|t-\Delta t} = x_{t-\Delta t|t-2\Delta t} + \eta_{t-\Delta t} \tag{3}$$

where $\eta_{t-\Delta t}$ is assumed uncorrelated in time, and is normally distributed with zero-mean and variance σ_η^2 (see section 2.2). The forecast error y_t (forecast minus observation at time *t*) is assumed corrupted from true forecast bias by a random error term ε_t:

$$y_t = x_t + \varepsilon_t = x_{t-\Delta t} + \eta_{t-\Delta t} + \varepsilon_t \tag{4}$$

where again ε_t is assumed uncorrelated in time and normally distributed with zero-mean and variance σ_ε^2 (see section 2.2). Thus, y_t includes systematic and random errors.

Kalman [1960] showed that the optimal recursive predictor of x_t (derived by minimizing Equation (1) with respect to β) can be written as a linear combination of the previous bias estimate and the previous forecast error:

$$\hat{x}_{t+\Delta t|t} = \hat{x}_{t|t-\Delta t} + \beta_{t|t-\Delta t}\left(y_t - \hat{x}_{t|t-\Delta t}\right) \tag{5}$$

where the hat (^) indicates the estimate.

To take into account the time-varying behavior of the bias that may occur at different times of the day, the filter algorithm is run on data for each hour of the day, using only values from previous days at the same hour. Moreover, if observations are missing for an hour, the filter uses the last known bias for that same hour from an earlier day. The true bias may change considerably in such a time period, and this creates spikes in the Kalman coefficients that can be smoothed by applying twice the following low-pass filter:

$$x_t = \frac{1}{2}\hat{x}_t + \frac{1}{4}\left[\hat{x}_{t-1} + \hat{x}_{t+1}\right] \tag{6}$$

To avoid negative forecast values, the Kalman-filtered ozone concentrations were truncated at a lower bound of 0 ppbv.

2.2 Variance Computation

The error variances σ_η^2 and σ_ε^2 are not usually known a priori. However, they can be estimated by defining the following new variable z_t as follows:

$$z_t = y_{t+\Delta t} - y_t = \eta_t + \varepsilon_{t+\Delta t} - \varepsilon_t \tag{7}$$

which was shown by *Dempster et al.* [1977] to have variance

$$\sigma_z^2 = \sigma_\eta^2 + 2\sigma_\varepsilon^2 \tag{8}$$

Define $r = \sigma_\eta^2 / \sigma_\varepsilon^2$, which can be substituted in (8) to give

$$\sigma_z^2 = r\sigma_\varepsilon^2 + 2\sigma_\varepsilon^2 = (2 + r)\sigma_\varepsilon^2 \tag{9}$$

The Kalman algorithm can be then used to estimate σ_ε^2 (which is a time-varying quantity). First, Equation (1) is applied:

$$p^{\sigma_\varepsilon^2}_{t|t-\Delta t} = \left(p^{\sigma_\varepsilon^2}_{t-\Delta t|t-2\Delta t} + \sigma^2_{\sigma_\eta^2}\right)\left(1 - \beta^{\sigma_\varepsilon^2}_{t|t-\Delta t}\right) \tag{10}$$

where $p^{\sigma_\varepsilon^2}$ is the expected mean-square-error in the σ_ε^2 estimate, $\sigma^2_{\sigma_\eta^2}$ is the variance of σ_η^2, and $\beta^{\sigma_\varepsilon^2}$ is the Kalman gain to estimate σ_ε^2.

Second, similarly to equation (2), the Kalman gain can be computed as follows:

$$\beta^{\sigma_\varepsilon^2}_{t+\Delta t|t} = \frac{p^{\sigma_\varepsilon^2}_{t|t-\Delta t} + \sigma^2_{\sigma_\eta^2}}{\left(p^{\sigma_\varepsilon^2}_{t|t-\Delta t} + \sigma^2_{\sigma_\eta^2} + \sigma^2_{\sigma_\varepsilon^2}\right)} \tag{11}$$

where $\sigma^2_{\sigma_\varepsilon^2}$ is the variance of σ_ε^2. Third, σ_ε^2 can be estimated by combining Equations (5) and (9):

$$\sigma^2_{\varepsilon,\, t+\Delta t|t} = \sigma^2_{\varepsilon,t|t-\Delta t} + \beta^{\sigma_\varepsilon^2}_{t|t-\Delta t}\left[\frac{\left(y_t - y_{t-\Delta t}\right)^2}{2+r} - \sigma^2_{\varepsilon,\, t|t-\Delta t}\right] \tag{12}$$

Constant values of 1 and 0.0005 are assigned to $\sigma^2_{\sigma_\varepsilon^2}$ and $\sigma^2_{\sigma_\eta^2}$, respectively [e.g., *Roeger et al.*, 2003].

Finally, σ_η^2 can be computed as $\sigma_\eta^2 = r\sigma_\varepsilon^2$. Then, the bias estimate ($\hat{x}$) can be computed by applying in sequence Equations (1), (2) and (5). This process is iterated through subsequent Δt.

3. Error-Ratio Sensitivity Tests

The KF performance is sensitive to the errors ratio $\sigma_\eta^2/\sigma_\varepsilon^2$. If the ratio is too high, the forecast-error white-noise variance (σ_ε^2) will be relatively small compared to the true forecast-bias white-noise variance (σ_η^2). Therefore, the filter will put excessive confidence on the previous forecasts, failing to estimate any forecast error. On the other hand, if the ratio is too low, the filter will be unable to respond to changes in bias. Consequently, there exists an optimal value for the ratio that is given by the climatology of the forecast region, which can be estimated by evaluating the filter performance in different situations with different meteorology and different AQ scenarios (not only for AQ episodes, as recognised by *Delle Monache et al.* [2005b]).

As described in section 1, the ICARTT data set offers a unique opportunity to test thoroughly the filter performance, both because of its length in time (56 days of summer 2004), and because includes eight different photochemical models, whose raw and KF predictions can be tested against surface observations from roughly 360 stations (for hourly ozone concentrations over the Northeast US and Southeast Canada, [*McKeen et al.*, 2005]). Specifically, with the ICARTT data set an optimal error-ratio value can be estimated, in order to produce a more accurate correction of ozone forecasts with the KF post-processing predictor method.

Delle Monache et al. [2005b] used a ratio value (0.01) from previous studies where the KF was used to bias-correct weather forecasts [*Roeger et al.*, 2003]. This value is close to the optimal value (0.06) found by *Homleid* [1995], who tested the filter for weather forecasts as well. Here the optimal ratio values (for ozone forecasts) are found by looking at the following statistical parameters:

- Pearson product-moment coefficient of linear correlation (herein "correlation"):

$$correlation = \frac{\sum_{i=1}^{N_{point}} \{[C_o(i) - \overline{C_o}][C_p(i) - \overline{C_p}]\}}{\sqrt{\sum_{i=1}^{N_{point}} [C_o(i) - \overline{C_o}]^2 \sum_{i=1}^{N_{hour}} [C_p(i) - \overline{C_p}]^2}} \quad (13)$$

- Root mean square error (RMSE):

$$RMSE = \sqrt{\frac{1}{N_{point}} \sum_{i=1}^{N_{point}} [C_p(i) - C_o(i)]^2} \quad (14)$$

where N_{point} is the number of all valid observation/prediction couples of 1-hour average concentrations over the 56-day period and 358 stations, $C_o(i)$ is the 1-hour average observed concentration at a monitoring station for hour t, $C_p(i)$ is the 1-hour average predicted concentration at a monitoring station for hour t, $\overline{C_o}$ is the average of 1-hour average observed concentrations over all the N_{point} observation/prediction couples available, $\overline{C_p}$ is the average of 1-hour average predicted concentrations over all the N_{point} observation/prediction couples available.

Correlation gives an indirect indication of the time lag between the predicted and measured ozone time series. The closer the correlation is to unity, the better is the correspondence of timing of ozone maximum and minimum between the two signals. RMSE gives important information about the skills of a forecast in predicting the magnitude of ozone concentration. It is also very helpful for understanding the filter performance, because it can be decomposed into systematic and unsystematic components (section 4.1.2).

Figures 2 and 3 show the correlation and RMSE values, respectively, for the eight models, with the ratio assuming values from 0.01 to 10 with increments of 0.01. Both

statistical parameters improve (i.e., higher correlation and lower RMSE) for all the models as the ratio value is increased from 0 to 0.5. Correlation values have their maxima and RMSE their minima roughly between 0.3 and 0.5. For ratio values greater than 0.5, the performance of all the models progressively deteriorates.

Based on the above results, in this study an optimal ratio value of 0.4 has been chosen. This optimal value is the result of an extensive sensitivity test performed on different models and several days of 1-hour predicted and observed ozone concentration at different locations (see section 1). Thus, this value can be the recommended value for future KF applications as a post-processing predictor bias-removal method for ozone forecasts.

4. Deterministic and Ensemble-Averaged Forecasts

In this section the filter performance is tested by evaluating the skills of 10 ozone forecasts and their KF corrected versions. These forecasts include the eight individual model forecasts, the ensemble-mean of the raw forecasts (E), and the ensemble mean of the KF forecasts (EK). Notably, when EK is filtered, the filter is applied twice (in combination with ensemble averaging) to the same signal. This double-filtered forecast has been found to have the best performance overall in *Delle Monache et al.* [2005b], and is tested here for comparison with that study.

The statistical metrics used for verification are correlation and RMSE as defined in section 3 with Equations (13) and (14), respectively, plus the following:

- unpaired peak prediction accuracy (UPPA):

$$UPPA = \frac{1}{N_{day} * N_{station}} \sum_{station=1}^{N_{station}} \left[\sum_{day=1}^{N_{day}} \frac{\left| C_p(day,station)_{\max} - C_o(day,station)_{\max} \right|}{C_o(day,station)_{\max}} \right] \quad (15)$$

- critical success index (CSI):

$$\frac{B}{A+B+C} \quad (16)$$

where N_{day} is the number of days, $N_{station}$ is the number of stations, $C_o(day,station)_{\max}$ is the maximum 1-hour average observed concentration at a monitoring station over one day, and $C_p(day,station)_{\max}$ is the maximum 1-hour average predicted concentration at a monitoring station over one day. CSI is computed for a given concentration threshold: A is the number of times the observation is below the threshold and the prediction is above it; B is the number of times both the observation and the prediction are above the

threshold; and *C* is the number of times the observation is above the threshold and the prediction is below it.

UPPA is included in the U.S. EPA guidelines [*U.S. Environmental Protection Agency (EPA)*, 1991] to analyze historical ozone episodes using photochemical grid models. The U.S. EPA acceptable-performance value is ± 20 %. UPPA is computed here as an average (over the days and stations available) of the absolute value of the normalized difference between the predicted and observed daily maximum at each station (equation (15)), so that under and over prediction are weighted equally and cancellation effects are not allowed. Thus, UPPA is non-negative and only the + 20 % acceptance performance upper limit is used in the next sections.

UPPA measures the ability of the forecasts to predict the ozone peak maximum on a given day. In the past, peak concentrations have been a primary concern for public health. However, in recent years over the midlatitudes of the Northern Hemisphere a rising trend for background ozone concentrations has been observed, while peak values are steadily decreasing [*Vingarzan*, 2004].

CSI has been chosen as a performance measure for forecasts of rare events because model and observed exceedances are equally weighted. It is compute here for thresholds between 60 and 90 ppbv, with increments of 2.5 ppbv.

4.1 Correlation

The closer correlation is to unity, the better. Figure 4 shows the results with this parameter for the eight model forecasts, E (their ensemble-averaged), and EK (the ensemble average of the filtered model forecasts). For each of these ten forecasts, the

black bar indicates the correlation of the raw forecast with the observations, while the white bar represents correlation for the Kalman filtered forecasts.

Among the raw deterministic forecasts (not including the ensemble-based forecast E and EK) WRF/CHEM-2 has the highest correlation. Kalman filtering provides significant improvements for AURAMS, BAMS-45, WRF/CHEM-1, WRF/CHEM-2, and STEM, going from 7 % (AURAMS and BAMS-45) to 20 % (STEM) higher correlation values. CHRONOS and EK correlations values are substantially the same after the correction (being slightly lower than the raw values). However, for BAMS-15, CMAQ/ETA, and E, after the correction correlation is worse after filtering, with E having the worst correction (-27 %). This is contrast with the results in *Delle Monache et al.* [2005b], where the correlation of the ensemble mean of the raw forecasts was improved after the correction, particularly at stations where the raw correlation values were low.

Nevertheless, E clearly has the highest correlation among the raw forecasts (as in *Delle Monache et al.* [2005b]) and EK has the highest value overall. The application of the filter twice (filtered EK) did not result in any improvement (contrary to *Delle Monache et al.* [2005b], as discussed further in the next sections), while ensemble averaging proves to ameliorate the correlation for both raw and filtered forecasts.

4.2 RMSE

The closer the values of this metric are to zero the better. RMSE is improved (lower values) for all the deterministic forecasts, except for BAMS-15 (Figure 5). E is worse after the correction, while EK has substantially the same RMSE before and after the Kalman correction, with its filtered version slightly worse. Among the raw forecasts

WRF/CHEM-2 has the lowest RMSE, while the best overall is again EK. The double filter application did not provide any improvement as found instead by *Delle Monache et al.* [2005b].

RMSE can be separated in different components. One decomposition was proposed by *Wilmott* [1981]. First, an estimate of concentration C^* is defined as follows:

$$C^*(i) = a + bC_o(i) \tag{17}$$

where a and b are the least-square regression coefficients of C_p and C_o (the predicted and observed ozone concentrations, respectively, as defined in section 3). Then the following two quantities can be defined:

$$RMSE_s = \sqrt{\frac{1}{N_{point}} \sum_{i=1}^{N_{point}} [C^*(i) - C_o(i)]^2} \tag{18}$$

$$RMSE_u = \sqrt{\frac{1}{N_{point}} \sum_{t=1}^{N_{point}} [C^*(i) - C_p(i)]^2} \tag{19}$$

where $RMSE_s$ is the RMSE systematic component, while $RMSE_u$ is the unsystematic one. $RMSE_s$ indicates the portion of error that depends on errors in the model, while $RMSE_u$ depends on random errors, on errors resulting by a model skill deficiency in predicting a specific situation, and on initial- and boundary-condition errors. The following is an interesting relationship between RMSE and its components:

$$RMSE^2 = RMSE_s^{\ 2} + RMSE_u^{\ 2} \tag{20}$$

The KF is expected to correct some of the systematic components of the errors (i.e., $RMSE_s$), while the unsystematic component ($RMSE_u$) on average (over the different

forecasts) should be affected little by the filter correction [*Delle Monache et al.,* 2005b]. In fact, if $RMSE_u$ is associated with errors introduced by model imperfections and initial-condition errors, then it cannot be removed except by fundamental model improvements or improvements in initial and boundary conditions.

Figure 6 shows the results for $RMSE_s$. There is an improvement for all the 10 forecasts after the KF correction, with $RMSE_s$ decreased from 12 (EK) to 82 % (STEM). Ensemble averaging does not reduce systematic error. The same kind of improvements for $RMSE_s$ has been found by *Delle Monache et al.* [2005b], even if less pronounced than what found in this study. The much greater duration of the data set used here and an optimal error-ratio value (as discussed in section 3) allow the filter to better capture the ozone-forecast systematic errors.

Unsystematic RMSE ($RMSE_u$) is never substantially improved with KF (Figure 7), and in few cases is even higher (for BAMS-15, CMAQ/ETA, E, and slightly also for EK) after the filter correction. Moreover, ensemble averaging reduces unsystematic error (filters out unpredictable components), confirming a finding by *Delle Monache et al.* [2005b].

4.3 UPPA

UPPA values closer to zero are better. BAMS-15 has the lowest UPPA among the raw forecasts (Figure 8). The UPPA values for the filtered forecasts are lower than for the raw versions of the same forecast, with improvements more pronounced than those presented *Delle Monache et al.* [2005b]. This statistical parameter confirms the benefits

of providing the filter with a much longer period to better learn the bias behavior, as well as the benefit of an optimal sigma error-ratio value.

UPPA Improvements range from 2 % (EK) to 48 % (STEM). The filtered EK clearly has the lowest (best) value. Along with E filtered and EK, those are the only forecasts having UPPA values clearly within the U.S. EPA acceptance value (20 %). This suggests the necessity of ensemble-averaging and Kalman filtering to accurately forecast ozone peak values that are the most harmful to our respiratory system.

4.4 CSI

CSI gives an indication of the forecast performance for rare events, and in this study is computed for ozone thresholds that span from 60 to 90 ppbv. During the ICARTT experiments, ozone above 60 ppbv was observed 6 % of the time, whereas ozone above 90 ppbv was observed 0.1 % of the time, out of a total of 421,082 valid observations. This means that the higher the threshold, the higher will be the sample uncertainty, and therefore the statistical significance gets progressively lower with higher thresholds.

The five-panel Figure 9 shows the results for AURAMS, BAMS-15, BAMS-45, CHRONOS and CMAQ/ETA, for the thresholds mentioned above, with increments of 2.5 ppbv. The continuous lines represent the raw forecast, and the dashed lines represent the Kalman-filtered forecasts. Figure 10 is similar to Figure 9, but for WRF/CHEM-1, WRF/CHEM-2, STEM, E, and EK.

The closer CSI is to 100 %, the better. The filter is improving the forecast performance with almost every threshold, except for CMAQ/ETA with 87.5 and 90 ppbv, and for EK with 90 ppbv. The largest improvements are observed with thresholds

between 60 and 75 ppbv, particularly for CHRONOS, STEM, WRF/CHEM-1, and E. Applying the filter twice (by filtering EK) does not produce any improvement. Namely, for rare events, the findings are similar to what was found with the other metrics (i.e., correlation, RMSE, and UPPA). The filter needs only one application to do its best correction. As already discussed, this reflects the benefits of having a long period to learn the bias behavior, as well as the use of an optimal error-ratio value.

The raw EK and the filtered E are always the better performing forecasts with the CSI metric, underlying the usefulness of ensemble averaging combined with Kalman filtering also to predict rare events. Among the raw deterministic forecasts, WRF/CHEM-2 has better skill than the others in predicting low frequency observed ozone concentration values.

5. Probabilistic Forecasts

The probability of an event occurrence (e.g., ozone concentration above a certain threshold), can be computed as the ratio of the number of the ensemble members that predict the event over the total number of members. The skill of a probabilistic forecast (PF) can be estimated by evaluating two attributes: resolution and reliability [*Jolliffe and Stephenson*, 2003]. In the following two subsections, these important attributes are defined and measured for a PF formed by the raw forecasts (PF-R), and a PF formed by the KF-corrected forecasts (PF-KF).

5.1 Resolution

Resolution measures the ability of the forecast to sort a priori the observed events into separate groups, when the events have a frequency different from the climatological frequency. A forecast with good resolution should be able to separate the observed events when two different probabilities are forecasted.

Resolution can be measured with Relative Operating Characteristics (ROC), developed in the field of signal-detection theory for discrimination between two alternative outcomes [*Mason*, 1982]. For the event to be forecasted, a contingency table is built (e.g., Table 2) for each forecast-probability threshold. With an eight-member ensemble, there are nine possible probability thresholds: from 0/8 to 8/8.

For Table 2, a count in forecast "YES" row means that the forecasted probability of the event (at the given time and station) is above the probability threshold, whereas a forecast "NO" results if it is below the threshold. For the columns in Table 2, an observation "YES" is counted if the observed concentration is above the fixed

concentration threshold (different from forecast-probability thresholds), and is "NO" otherwise. Once *I*, *II*, *III* and *IV* (combinations of forecast and observation "YES" and "NO" counts, as in Table 2) are detected for all the times and stations available, the *hit rate* and *false-alarm rate* can be computed for a given forecast-probability threshold as follows:

$$hit\ rate = \frac{I}{I + III} = \frac{number\ of\ event\ correct\ forecasts}{total\ number\ of\ event\ occurrences} \tag{21}$$

$$false\text{-}alarm\ rate = \frac{II}{II + IV} = \frac{number\ of\ event\ non\ correct\ forecasts}{total\ number\ of\ event\ non\ occurrences} \tag{22}$$

After the *hit rate* and *false alarm rate* are computed for each of the nine possible forecast-probability thresholds, *hit rates* can be plotted on the ordinate against the corresponding *false-alarm rates* on the abscissa to generate the ROC curve. Figure 11 shows the ROC curve for PF-R, where the event to be forecasted is ozone above 50 ppbv, and the labels adjacent to the asterisks indicate the forecast-probability threshold relative to that point.

For a PF with good resolution, the ROC curve is close to the upper left hand corner of the graph. The area under the ROC curve (shaded in Figure 11) quantifies the ability of an ensemble to discriminate between events, which can be equated to forecast usefulness, and is known also as the ROC score [*Mason and Graham*, 1999]. The closer the area is to unity, the more useful is the forecast. A value of 0.5 indicates that the forecast system has no skill (area below the dashed line in Figure 11), relative to a chance forecast from climatology. The ROC curve does not depend on the forecast bias, hence is independent of reliability (section 5.2). The ROC represents a PF intrinsic value.

Figure 12 shows the ROC-area values for PF-R and PF-KF. The ROC area is

computed for ozone concentration thresholds from 10 to 90 ppbv, with increments of 10 ppbv. Kalman filtering is able to improve considerably the PF predictive skill between 10 and 50 ppbv. However, from 70 to 90 ppbv it degrades the PF resolution, even though PF-KF ROC-Area values are still above 0.85 with these two thresholds, indicating a forecast with high resolution. This means that the filter is not only removing the bias, but it is also modifying the predictive skill of the forecasts, by improving those below 60 ppbv, and deteriorating those above it. Resolution is not affected by removing the overall bias by definition, but since here KF is applied for each hour of the day independently (section 2.1), it predicts different biases for different hours, and then is also able to affect the forecast resolution.

5.2 Reliability

Reliability measures the capability of a PF to predict unbiased estimates of the observed frequency associated with different forecast probabilities. In a perfectly reliable forecast, the forecasted probability of the event should be equal to the observed frequency of the event for all the cases when that specific probability value is forecasted. Reliability alone is not sufficient to establish if a PF produces valuable forecasts or not. For instance, a system that always forecasts the climatological probability of an event is reliable, but not useful.

Reliability can be measured with a Talagrand diagram [*Talagrand and Vautard*, 1997], also known as the rank histogram [*Hamill and Colucci*, 1997]. First, the ensemble members are ranked for each prediction. Then, the frequency of an event occurrence in each bin of the rank histogram is computed and plotted against the bins. The number of

bins equals the number of ensemble members plus one. A perfectly reliable PF shows a flat Talagrand diagram, where the bins have all the same height ("ideal bin height"). In fact, if each ensemble member represents an equally likely evolution of the ozone concentration, the observations are equally likely to fall between any two members.

Figure 13 shows the Talagrand diagram for PF-R (black bars) and PF-KF (white bars). The PF-R forecast is positively biased, because the highest frequency is reported on the first bin and the frequency generally decreases with increasing bin number. This means that the observations, when ranked with the observation at a given time and station, tend to fall more often in the lower bins, indicating over prediction.

The PF-KF Talagrand diagram is much closer to the ideal flat shape (indicated by the continuous line), meaning that the filter is removing successfully the bias from the individual forecasts, and this in turn results in a much more reliable probabilistic prediction.

6. Summary and Conclusions

This study presents an in-depth analysis of the Kalman filter (KF) as post-processing predictor bias-correction method for deterministic, ensemble-averaged, and probabilistic surface ozone forecasts. The skills of raw and Kalman-filtered ozone forecasts have been evaluated against observations collected during the summer 2004 in the Northeast US and Southeast Canada, as part of the Intercontinental Transport and Chemical Transformation (ICARTT) New England Air Quality (AQ) Study [*McKeen et al.*, 2005]. The completeness of this data set, including 1-hour ozone forecasts from eight different models for 56 days, and observations from roughly 360 stations, offered a unique opportunity to test thoroughly the filter performance.

An optimal KF error-ratio parameter value of 0.4 has been found by evaluating the filter performance in different situations with different meteorology and different AQ scenarios. This value is recommended for future KF applications as a post-processing predictor bias-removal method for ozone forecasts.

The correlation results show that Kalman filtering reduces significantly the time lag between the predicted and measured ozone time series for all the forecasts (except for the Baron Advanced Meteorological System Multi-scale Air Quality Simulation Platform run at 15 km (BAMS-15), The Community Multi-scale Air Quality Model (CMAQ/ETA), and the ensemble mean of the raw forecasts E). The forecast having the least lag is the ensemble average of the Kalman filtered forecasts (EK), while for raw deterministic forecasts the Weather Research and Forecast Model/Chemistry model version 2.03 (WRF/CHEM-2) has the least lag. Ensemble averaging increases the correlation with the observations for both raw and filtered forecasts.

For all the deterministic forecasts (except for BAMS-15), the KF improves the ability to predict the ozone-concentration magnitude (based on the root mean square error metric, RMSE). Among the raw forecasts WRF/CHEM-2 has the lowest RMSE, while the best RMSE overall is again for EK. The tests involving RMSE systematic ($RMSE_S$) and unsystematic ($RMSE_u$) components confirmed the results from *Delle Monache et al.* [2005b]: the filter removes a good portion of the bias while it has a minimal affects on the random errors. Vice versa, ensemble averaging tends to remove the unsystematic component of RMSE, while it leaves substantially unaltered the bias. For this reason (considering also the other statistical metrics), the combination of Kalman filtering and ensemble averaging (i.e., EK) resulted in the best forecasts in this study.

KF improves the ability to predict the daily surface surface ozone maximum concentration magnitude. Comparing the Unpaired Peak Prediction Accuracy (UPPA) metric results, The filtered EK has the lowest (best) value, while BAMS-15 has the lowest UPPA among the raw deterministic forecasts. Along with E filtered and EK, those are the only forecasts having UPPA values within the U.S. EPA acceptance value (20 %, *U.S. Environmental Protection Agency(EPA)* [1991]). This suggests the necessity of ensemble-averaging and Kalman filtering to accurately forecast the surface ozone peak magnitude.

After the forecasts are Kalman filtered, their ability to predict rare events is improved consistently, giving higher Critical Success Index (CSI) values for almost all the concentration thresholds considered in this study. EK and the filtered E are always better than the other forecasts in forecasting these low-frequency events, demonstrating also in

these cases the usefulness of ensemble averaging combined with Kalman filtering. WRF/CHEM-2 has the highest CSI values among the raw deterministic forecasts.

Kalman filtering is able to improve considerably the probabilistic-forecast (PF) predictive skill for ozone concentrations above thresholds from 10 to 50 ppbv. However, from 70 to 90 ppbv it degrades the PF resolution, even though the ROC-Area values are still above 0.85 with these two thresholds, indicating a forecast with high resolution. Thus, the filter is not only removing the bias, but it is also modifying the predictive skill of the forecast, by improving them below 60 ppbv, and deteriorating them above it. Resolution is not affected by removing the overall bias, by definition, but since here KF is applied for each hour of the day independently, it predicts different biases for different hours, and is thus able to affect the forecast resolution.

The Talagrand diagrams show that the PF composed by raw forecasts is positively biased, whereas PF including the Kalman filtered forecasts is much closer to the ideal flat shape, meaning that the filter removes successfully most of the bias from the individual forecasts, and this in turn results in a much more reliable probabilistic prediction.

Finally, the results of this study indicates that only one application of the Kalman filter is needed to achieve the best correction (compared to earlier findings by *Delle Monache et al.* [2005b] suggesting that two applications of the filter are useful). This reflects the benefits of having a longer period to learn the bias behavior (as with the ICARTT data set used here), as well as the use of an optimal error-ratio value.

The significance of Kalman-filter postprocessing and ensemble averaging is that they are both effective for real-time AQ forecasting. Namely, they reduce both systematic biases and random errors from coupled meteorological and chemistry transport models to

give the best estimate of future conditions, regardless of the synoptic situation and for AQ scenarios for which the underlying models were not specifically tuned.

This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

References

Atkinson, R., D. L. Baulch, R. A. Cox, R. F. Hampson, J. A. Kerr, and J. Troe (1992), Evaluated kinetic and photochemical data for atmospheric chemistry: supplement IV, *Atmos. Environ.*, 26A, 1187-1230.

Binkowski, F. S., and U. Shankar (1995), The regional particulate model. I. Model description and preliminary results, *J. Geophys. Res.*, 100, D12, 26191–26209.

Bozic, S. M. (1994), *Digital and Kalman Filtering*, 2nd ed., 160 pp., John Wiley & Sons, New York.

Burgers, G., P. J. van Leeuwen, and G. Evensen (1998), Analysis scheme in the ensemble Kalman filter, *Mon. Weather Rev.*, 126, 1719–1724.

Byun, D. W., and J. K. S. Ching (Eds.) (1999), Science algorithms of the EPA Models-3 Community Multiscale Air quality (CMAQ) modeling system, *EPA/600/R-99/030*, Off. of Res. and Dev., U.S. Environ. Prot. Agency, Washington, D. C..

Cannon, A. J. and, E. R. Lord (2000), Forecasting summertime surface level ozone concentrations in the Lower Fraser Valley of British Columbia: An ensemble neural network approach, *Journal of the Air & Waste Management Association*, 50, 322-339.

Carmichael, G. R., and Coauthors (2003), Regional-scale chemical transport modeling in support of intensive field experiments: overview and analysis of the TRACE-P observations, *J. Geophys. Res.,* 10.1029/2002JD003117.

Carter, W. (2000), Documentation of the SAPRC-99 chemical mechanism for VOC reactivity assessment, *Final Report to California Air Resources Board Contract No. 92-329*, University of California Riverside, Riverside, CA.

Coats, C.J. Jr. (1996), High-performance algorithms in the Sparse Matrix Operator Kernel Emissions (SMOKE) modeling system, Proc. 9th AMS Joint Conference on Applications of Air Pollution Meteorology with A&WMA, Amer. Meteor. Soc., Atlanta, GA, 584-588.

Côté, J., S. Gravel, A. Méthot, A. Patoine, M. Roch and A. Staniforth (1998a), The operational CMC-MRB Global Environmental Multiscale (GEM) model. Part I: Design considerations and formulation, *Mon. Wea. Rev.*, 126, 1373-1395.

—, J., J.-G. Desmarais, S. Gravel, A. Méthot, A. Patoine, M. Roch and A. Staniforth (1998b), The operational CMC/MRB Global Environmental Multiscale (GEM) model. Part II: Results, *Mon. Wea. Rev.*, 126, 1397-1418.

Delle Monache, L., and R. Stull (2003), An ensemble air quality forecast over western Europe during an ozone episode, *Atmos. Environ.*, 37, 3469–3474.

—, X. Deng, Y. Zhou, and R. Stull (2005a), Ozone ensemble forecasts. Part I: a new ensemble design, *J. Geophys. Res.*, 111, D05307, doi:10.1029/2005JD006310.

—, T. Nipen, X. Deng, Y. Zhou, and R. Stull (2005b), Ozone ensemble forecasts. Part II: a Kalman-filter predictor bias correction, *J. Geophys. Res.*, 111, D05308, doi:10.1029/2005JD006311.

Dempster, A., N. Laird, and D. Rubin (1997), Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society*, 39, 1-38.

Gery, M. W., and Coauthors (1989), A photochemical kinetics mechanism for urban and regional scale computer models, *J. Geophys. Res.*, 94, 12295-12356.

Grell, G. A., J. Dudhia, and D. R. Stauffer (1994), A description of the fifth-generation Penn State/NCAR Mesoscale Model (MM5), NCAR Tech. Note, *NCAR/TN-398+STR*, 122 pp.

—, G. A., S. E. Peckham, R. Schmitz, S. A. McKeen, G. Frost, W. Skamarock, and B. Eder (2005), Fully coupled "online" chemistry within the WRF model, *Atmos. Environ.*, 39, 6957-6975.

Guenther, A. B., and Coauthors (1995), A global model of natural volatile organic compound emissions, *J. Geophys. Res.*, 100D, 8873-8892.

Hamill, T. M., and S. J. Colucci (1997), Verification of Eta–RSM short-range ensemble forecasts, *Mon. Wea. Rev.,* 125, 1312–1327.

—, and C. Snyder (2000), A hybrid ensemble Kalman filter-3D variational analysis scheme, *Mon. Weather Rev.*, 128, 2905–2919.

Homleid, M. (1995), Diurnal corrections of short-term surface temperature forecasts using Kalman filter, *Weather and Forecasting*, 10, 989-707.

Houtekamer, P. L., and H. L. Mitchell (2001), A sequential ensemble Kalman filter for atmospheric data assimilation, *Mon. Weather Rev.*, 129, 123–137.

—, H. L. Mitchell, G. Pellerin, M. Buehner, M. Charron, L. Spacek, and B. Hansen (2005), Atmospheric data assimilation with an ensemble Kalman filter: results with real observations, *Mon. Weather Rev*., 133, 604-620.

Jolliffe, I. T., and D. B. Stephenson (2003), *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, 240 pp., Wiley and Sons, West Sussex, United Kingdom.

Kalman, R. E. (1960), A new approach to linear filtering and prediction problems, *Journal of Basic Engineering*, 82, 35-45.

Lurman, F. W., A. C. Lloyd, and R. Atkinson (1986), A chemical mechanism for use in long-range transport/acid deposition computer modeling, *J. Geophys. Res.*, 91, 10905-10936.

Mason, I. (1982), A model for assessment of weather forecasts, *Aust. Meteor. Mag.,* 30, 291–303.

Mason, S. J., and N. E. Graham (1999), Conditional probabilities, relative operating characteristics, and relative operative levels, *Weather and Forecas*ting, 14, 713-725.

McHenry, J. N., W. F. Ryan, N. L. Seaman, C. J. Coats Jr., J. Pudykiewicz, S. Arunachalam, and J. M. Vukovich (2004), A real-time Eulerian photochemical model forecast system, *Bull. Amer. Meteor. Soc.*, 85, 525–548.

McKeen, S. A., G. Wotawa, D. D. Parrish, J. S. Holloway, M. P. Buhr, G. Hubler, F. C. Fehsenfeld, and J. F. Meagher (2002), Ozone production from Canadian wildfires during June and July of 1995, *J. Geophys. Res.*, 107(D14), doi:10.1029/2001 JD000697.

—, S. A., J. M. Wilczak, G. A. Grell, I. Djalalova, S. Peckham, E.-Y. Hsie, W. Gong, V. Bouchet, S. Menard, R. Moffet, J. McHenry, J. McQueen, Y. Tang, G. R. Carmichael, M. Pagowski, A. Chan, T. Dye, G. Frost, P. Lee, and R. Mathur (2005), Assessment of an ensemble of seven real-time ozone forecasts over Eastern North America during the summer of 2004, *J. Geophys. Res.*, 110, D21307, doi:10.1029/2005JD005858.

McQueen, J., and Coauthors (2004), J2.16, Development and evaluation of the NOAA/EPA prototype air quality model prediction system, Preprints, 20th

Conference on Weather Analysis and Forecasting/16th Conference on Numerical Weather Prediction, Seattle, WA.

Moran, M. D., M. T. Scholtz, C. F. Slama, A. Dorkalam, A. Taylor, N. S. Ting, D. Davies, P. A. Makar, and S. Venkatesh (1997), An overview of CEPS1.0: Version 1.0 of the Canadian Emissions Processing System for regional-scale air quality models, Proc. 7th AWMA Emission Inventory Symp., Research Triangle Park, North Carolina, Air & Waste Management Association, Pittsburgh.

Pagowski, M., G. A. Grell, S. A. McKeen, D. Dévényi, J. M. Wilczak, V. Bouchet, W. Gong, J. McHenry, S. Peckham, J. McQueen, R. Moffet, and Y. Tang (2005a), A simple method to improve ensemble-based ozone forecasts, *Geophys. Res. Lett.*, 32, L07814, doi:10.1029/2004GL022305.

—, M., and Coauthors (2005b), Application of dynamic linear regression to improve skill of ensemble-based deterministic ozone forecasts, accepted to appear on *Atmos. Environ.*.

Pudykiewicz, J., A. Kallaur, and P. K. Smolarkiewicz (1997), Semi-Lagrangian modeling of tropospheric ozone, *Tellus*, 49B, 231-258.

Roeger, C., R. B. Stull, D. McClung, J. Hacker, X. Deng, and H. Modzelewski (2003), Verification of mesoscale numerical weather forecast in mountainous terrain for application to avalanche prediction, *Weather and Forecasting*, 18, 1140-1160.

Russell, A., and R. Dennis (2000), NARSTO critical review of photochemical models and modeling, *Atmos. Environ.*, 34, 2283-2324.

Segers, A. J., H. J. Eskes, R. J. van der A, R. F. van Oss, P. F. J. van Velthoven (2005), Assimilation of GOME ozone profiles and a global chemistry-transport model

using a Kalman filter with anisotropic covariance, *Quart. J. Roy. Meteor. Soc.*, 131, 477-502.

Stockwell, W. R., P. Middleton, J. S. Chang, and X. Tang (1995), The effect of acetyl peroxy-pperoxy radical reactions on peroxyacetyl nitrate and ozone concentrations, *Atmos. Environ*., 29, 1591-1599.

Talagrand, O., and R. Vautard (1997), Evaluation of probabilistic prediction systems, Proceedings ECMWF Workshop on Predictability*,* ECMWF, Reading, United Kingdom, 1–25.

U.S. Environmental Protection Agency, (1991), Guideline for Regulatory Application of the Urban Airshed Model, *USEPA Rep. EPA-450/4-91-013*, Research Triangle Park, N. C..

van Loon, M., P. J. H. Builtjes, A. J. Segers (2000), Data assimilation applied to LOTOS: first experiences, *Environ. Mod. Softw.*, 15, 603-609.

Vingarzan, R. (2004), A review of surface ozone background levels and trends, *Atmos. Environ.*, 38, 3431-3442.

Wilczak, J. M., S. A. McKeen, I. Djalalova, and G. Grell (2005), Bias-corrected ensemble predictions of surface O_3, submitted to the *J. Geophys. Res*.

Willmott, C. J. (1981), On the validation of models, *Phys. Geogr.*, 2, 184-194.

Figure Captions

Figure 1. Eight photochemical model domains (left) and their overlap, including stations subdivided by urban, suburban, rural and unknown classification (right).

Figure 2. Correlation values (Equation 13) for the eight photochemical models, computed with sigma error-ratio values ranging from 0.01 to 10, with increments of 0.01. Values are within the interval [-1, 1], with correlation = 1 being the best possible value.

Figure 3. Similarly to Figure 2, but for root mean square error (RMSE) (ppbv) (Equation 14). Values are within the interval [0, +∞), with a perfect forecast when RMSE = 0.

Figure 4. Correlation values (Equation 13) for the eight models, the ensemble mean of the raw forecasts (E), and the ensemble mean of the Kalman filtered forecasts (EK). Black bars represent the raw forecasts, and white bars indicate the values for the Kalman filtered forecasts. Values are within the interval [-1, 1], with correlation = 1 being the best possible value.

Figure 5. Similarly to Figure 4, but for root mean square error (RMSE) (ppbv) (Equation 14). Values are within the interval [0, +∞), with a perfect forecast when RMSE = 0.

Figure 6. Similarly to Figure 4, but for the root mean square error (RMSE) systematic component (ppbv) (Equation 18).

Figure 7. Similarly to Figure 4, but for the root mean square error (RMSE) unsystematic component (ppbv) (Equation 19).

Figure 8. Similarly to Figure 4, but for the unpaired peak prediction accuracy (UPPA) (%) (Equation 15). The continuous lines are the U.S. EPA acceptance values (+ 20 %). Values are within the interval [0, +∞), with a perfect peak forecast when UPPA = 0.

Figure 9. Critical success index (CSI) (%) values (Equation 16) for (from top to the bottom panel) AURAMS, BAMS-15, BAMS-45, CHRONOS and CMAQ/ETA. CSI is compute for ozone above thresholds ranging from 60 to 90 ppbv, with increments of 2.5 ppbv. Values are within the interval [0, 100], with a perfect peak forecast when CSI = 100.

Figure 10. Similar to Figure 9, but for (from top to the bottom panel) WRF/CHEM-1, WRF/CHEM-2, STEM, the ensemble mean of the raw forecasts (E) and of the Kalman filtered forecast (EK).

Figure 11. Relative Operating Characteristics (ROC) curve for the probabilistic forecast formed by the raw forecasts (PF-R), for observed ozone concentration above 50 ppbv. The better the probabilistic forecast, the closer the ROC curve is to the upper left corner. The shaded portion of the plot represents the ROC area (larger areas are better), and the dashed line is the ROC curve for a chance forecast. Hit rates are plotted on the ordinate against the corresponding false-alarm rates on the abscissa, to generate the ROC curve for each probability threshold (the labels adjacent to the asterisks), where the probability threshold assumes values from 0/8 to 8/8, with increments of 1/8.

Figure 12. Relative Operating Characteristics (ROC) area values for different ozone concentration thresholds (ranging from 10 to 90 ppbv, with increments of 10

ppbv), for the probabilistic forecast formed by the raw forecasts (PF-R) (continuous line) and the probabilistic forecast formed by the Kalman-filtered corrected forecasts (PF-KF) (dashed line). Values are within the interval [0, 1], with the perfect ROC-area = 1.

Figure 13. Talagrand diagram (rank histogram) for the probabilistic forecast formed by the raw forecasts (PF-R) (black bars) and the probabilistic forecast formed by the Kalman-filtered corrected forecasts (PF-KF) (white bars). The number of bins equals the number of ensemble members plus one. The solid horizontal line represents the perfect Talagrand diagram shape (flat). The closer the diagram to this horizontal line, the better.

Table 1. General information about the eight photochemical models used in this study.

	Driving Meteorology	Chemical Mechanism	Emissions	Horizontal Spatial Resolution (km)
AURAMS	GEM [*Côté et al.*, 1998a,b]	ADOM II [*Lurman*, 1986; *Atkinson et al.*, 1992]	CEPS [*Moran et al.*, 1997]	42
CHRONOS	GEM [*Côté et al.*, 1998a,b]	ADOM II [*Lurman*, 1986; *Atkinson et al.*, 1992]	CEPS [*Moran et al.*, 1997]	21
BAMS-15	MM5 [*Grell et al.*, 1994]	CB-IV [*Gery et al.*, 1989]	SMOKE [*Coats*, 1996]	15
BAMS-45	MM5 [*Grell et al.*, 1994]	CB-IV [*Gery et al.*, 1989]	SMOKE [*Coats*, 1996]	15
CMAQ/ETA	NWS/NCEP ETA [*McQuenn et al.*, 2004]	CB-IV [*Binkowski and Shankar*, 1995]	SMOKE [*Coats*, 1996]	12
WRF/CHEM-1	WRF [*Grell et al.*, 2005]	RADM2 [*Stockwell et al.*, 1995]	[*McKeen et al.*, 2002]	27
WRF/CHEM-2	WRF [*Grell et al.*, 2005]	RADM2 [*Stockwell et al.*, 1995]	[*McKeen et al.*, 2002] Improved emission inventory with respect to WRF/CHEM-1	27
STEM	MM5 [*Grell et al.*, 1994]	SAPRC-99 [*Carter*, 2000]	U.S. EPA NEI-99 and IGAC-GEIA archive [*Guenther et al.*, 1995]	12

Table 2. Contingency table to compute *hit rates* and *false-alarm rates* for a given event and a forecast-probability threshold.

Contingency Table		***Observation***	
		YES	**NO**
Forecast	**YES**	***I***	***II***
	NO	***III***	***IV***

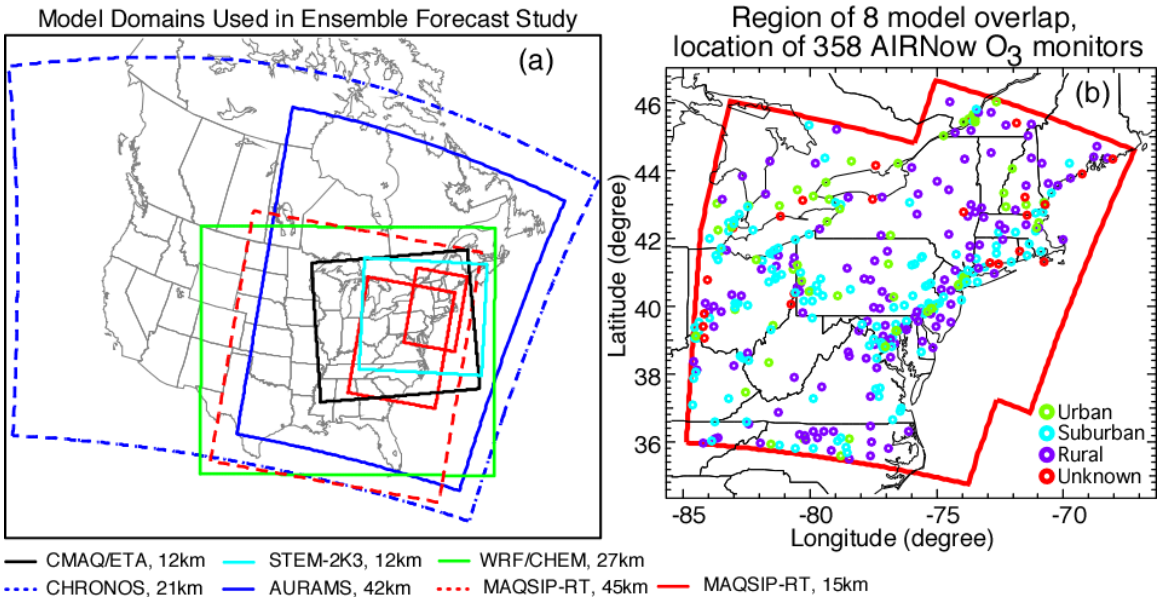

Figure 1. Eight photochemical model domains (left) and their overlap, including stations subdivided by urban, suburban, rural and unknown classification (right).

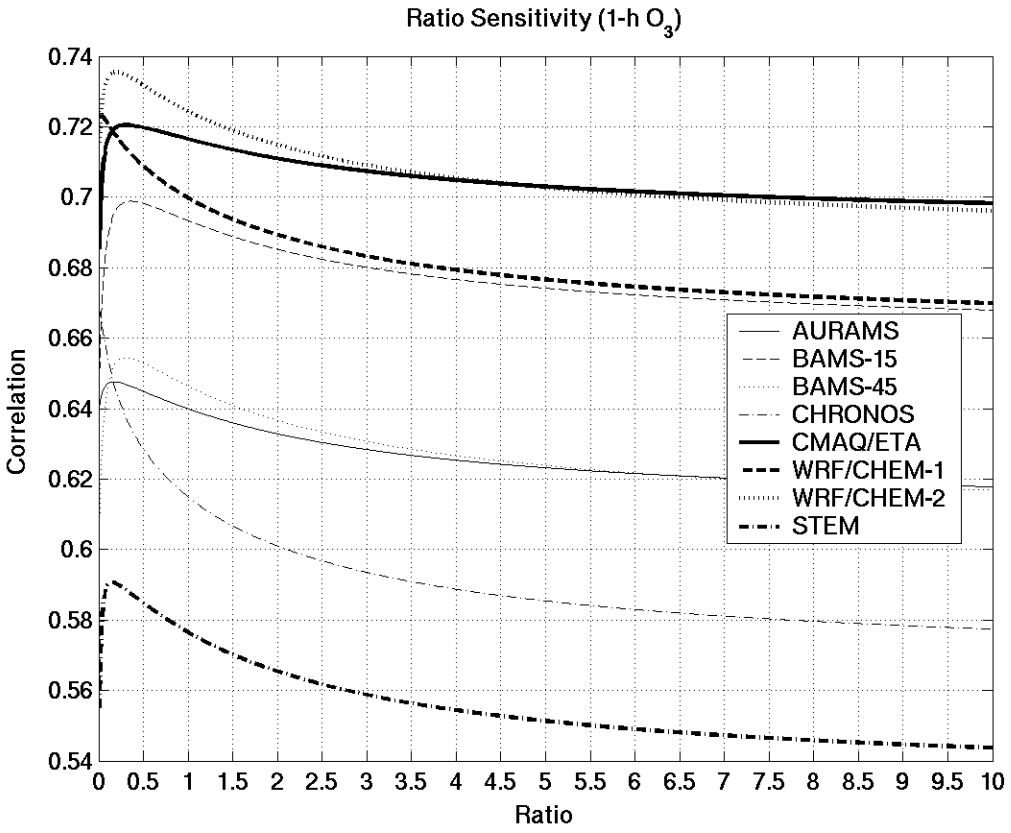

Figure 2. Correlation values (Equation 13) for the eight photochemical models, computed with sigma error-ratio values ranging from 0.01 to 10, with increments of 0.01. Values are within the interval [-1, 1], with correlation = 1 being the best possible value.

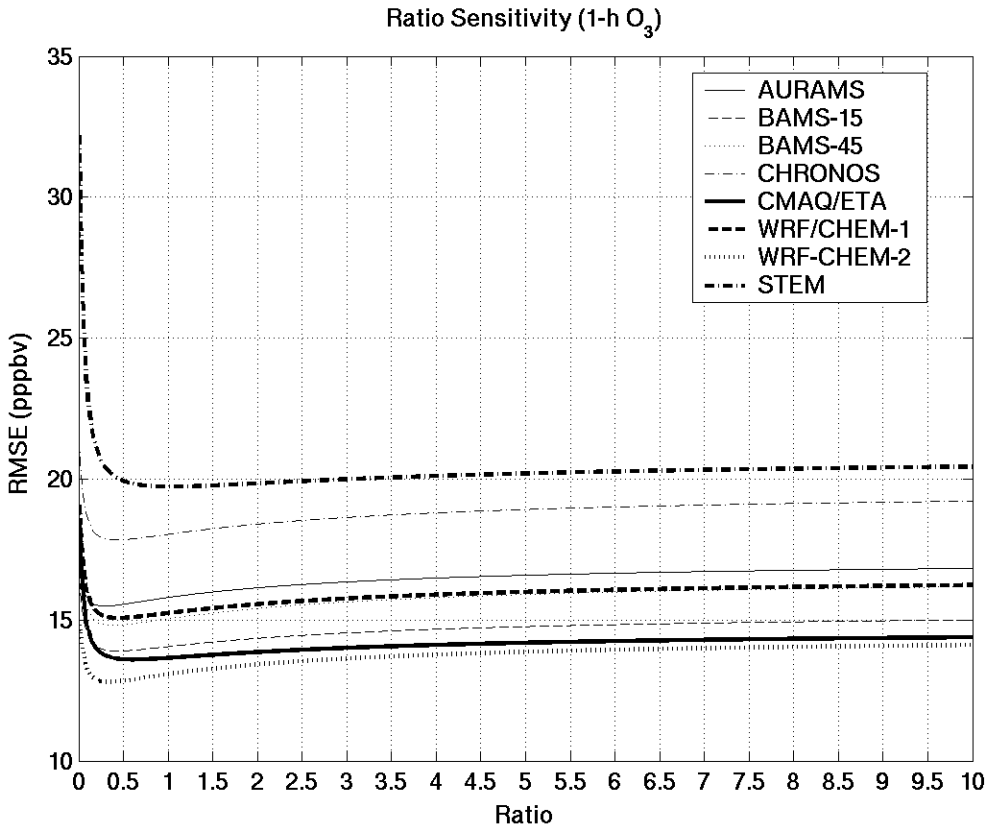

Figure 3. Similarly to Figure 2, but for root mean square error (RMSE) (ppbv) (Equation 14). Values are within the interval [0, +∞), with a perfect forecast when RMSE = 0.

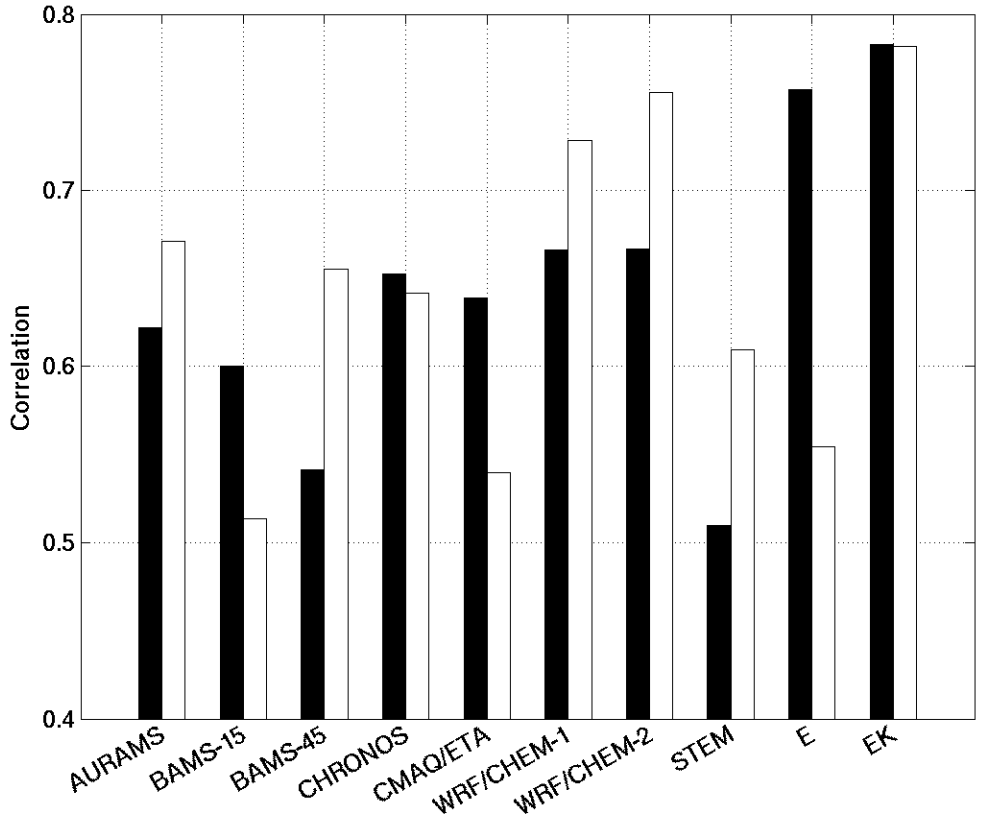

Figure 4. Correlation values (Equation 13) for the eight models, the ensemble mean of the raw forecasts (E), and the ensemble mean of the Kalman filtered forecasts (EK). Black bars represent the raw forecasts, and white bars indicate the values for the Kalman filtered forecasts. Values are within the interval [-1, 1], with correlation = 1 being the best possible value.

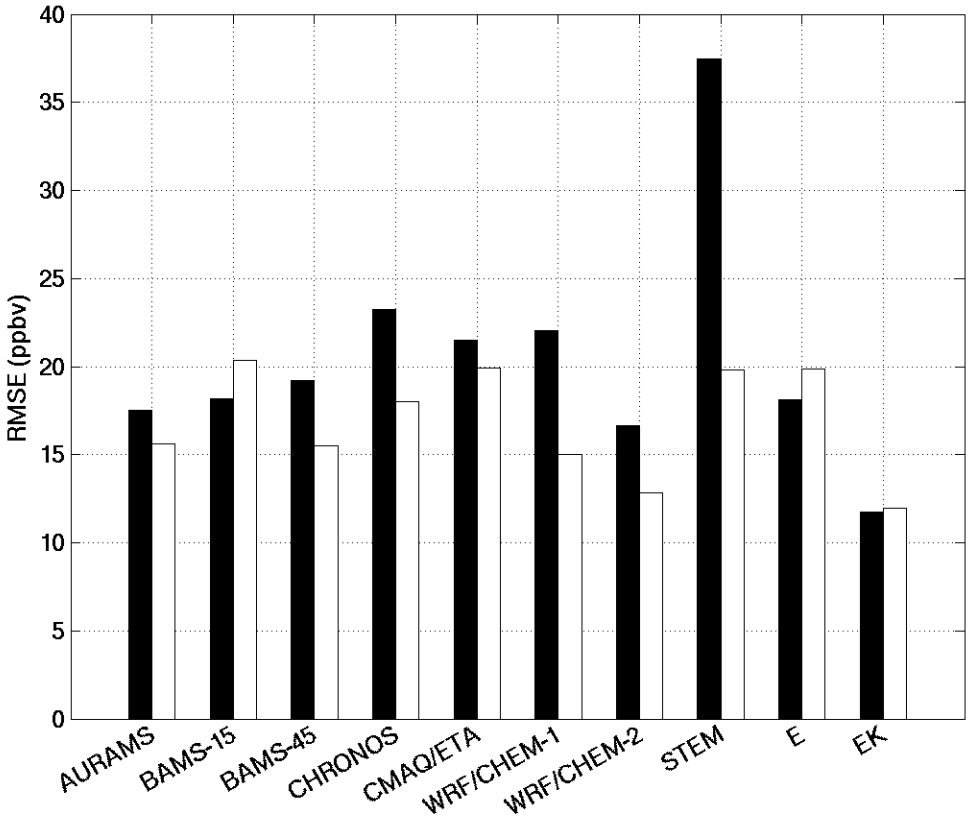

Figure 5. Similarly to Figure 4, but for root mean square error (RMSE) (ppbv) (Equation 14). Values are within the interval [0, +∞), with a perfect forecast when RMSE = 0.

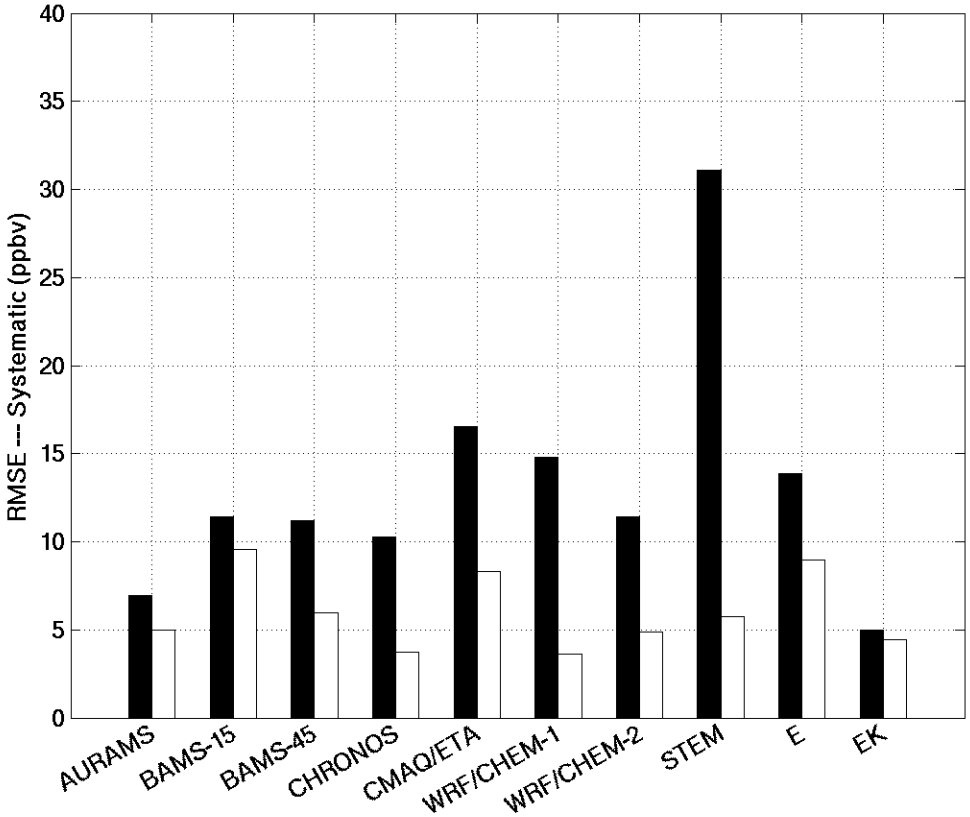

Figure 6. Similarly to Figure 4, but for root mean square error (RMSE) systematic component (ppbv) (Equation 18).

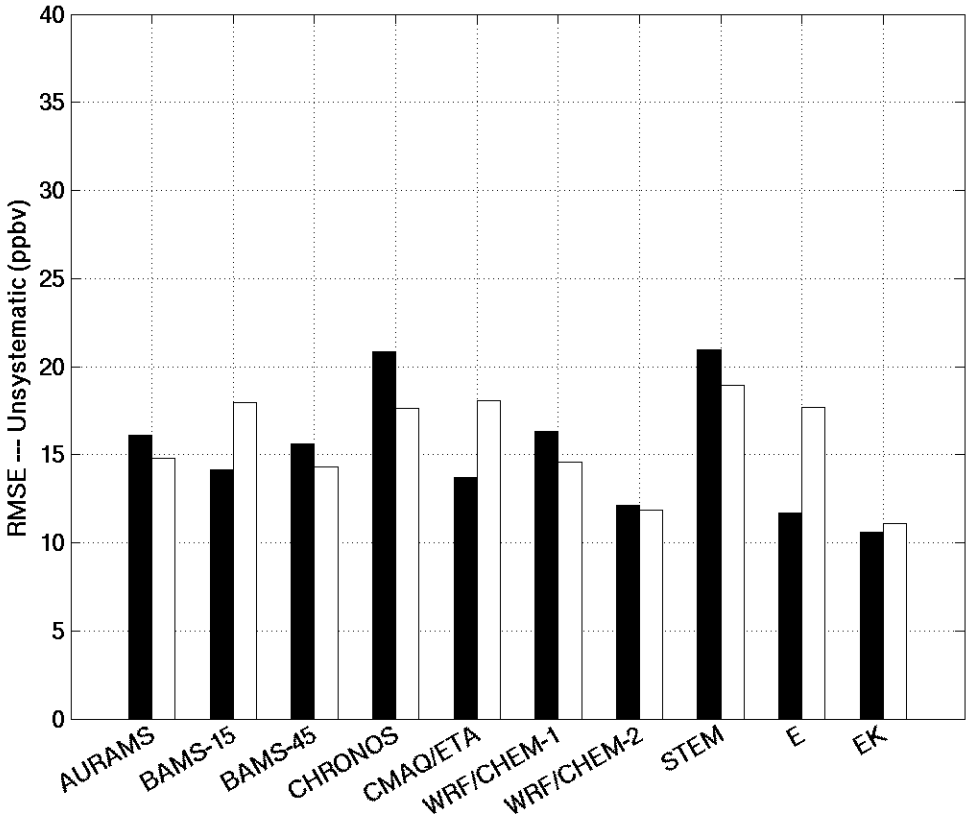

Figure 7. Similarly to Figure 4, but for root mean square error (RMSE) unsystematic component (ppbv) (Equation 19).

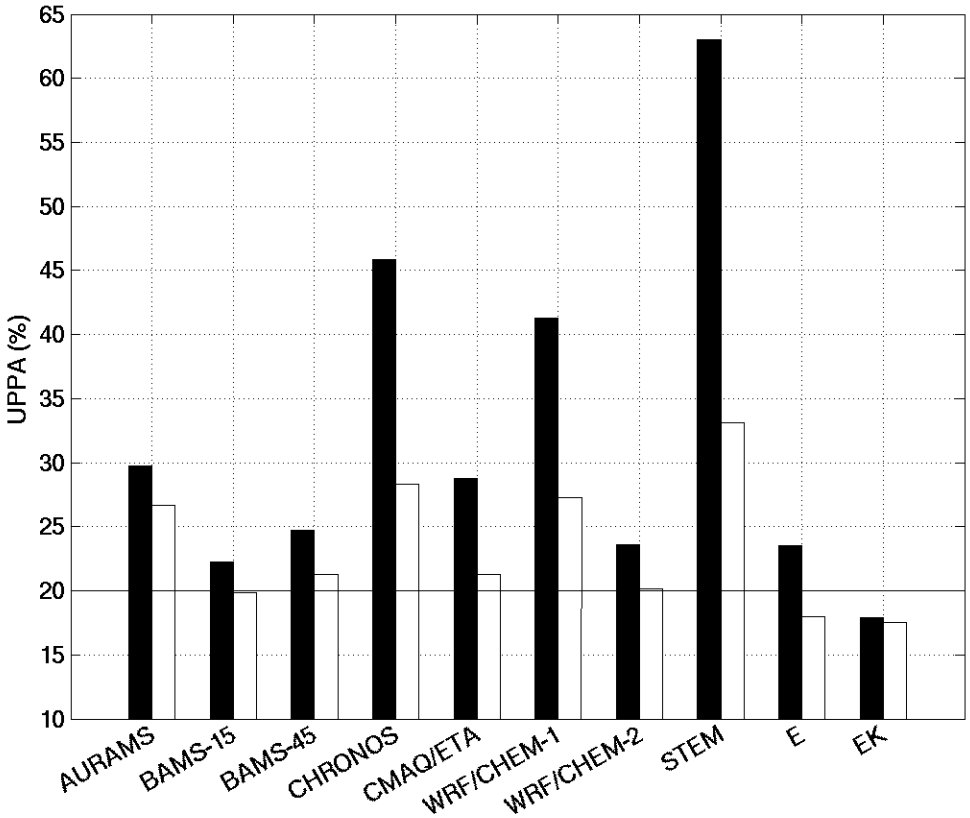

Figure 8. Similarly to Figure 4, but for the unpaired peak prediction accuracy (UPPA) (%) (Equation 15). The continuous lines are the U.S. EPA acceptance values (+ 20 %). Values are within the interval [0, +∞), with a perfect peak forecast when UPPA = 0.

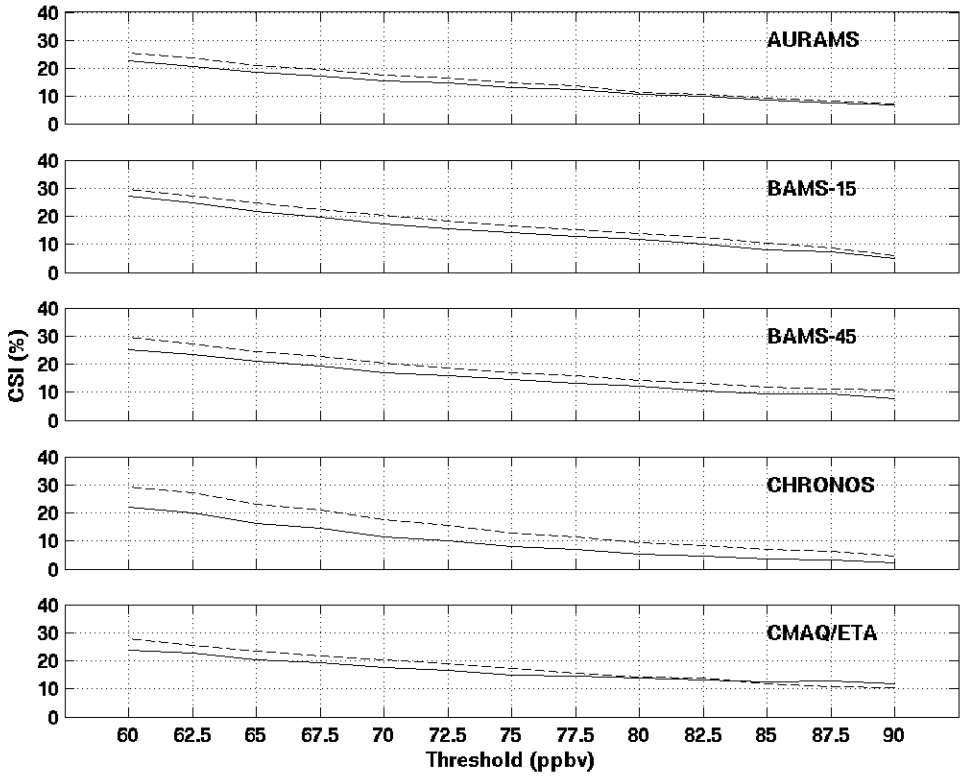

Figure 9. Critical success index (CSI) (%) values (Equation 16) for (from top to the bottom panel) AURAMS, BAMS-15, BAMS-45, CHRONOS and CMAQ/ETA. CSI is compute for ozone above thresholds ranging from 60 to 90 ppbv, with increments of 2.5 ppbv. Values are within the interval [0, 100], with a perfect peak forecast when CSI = 100.

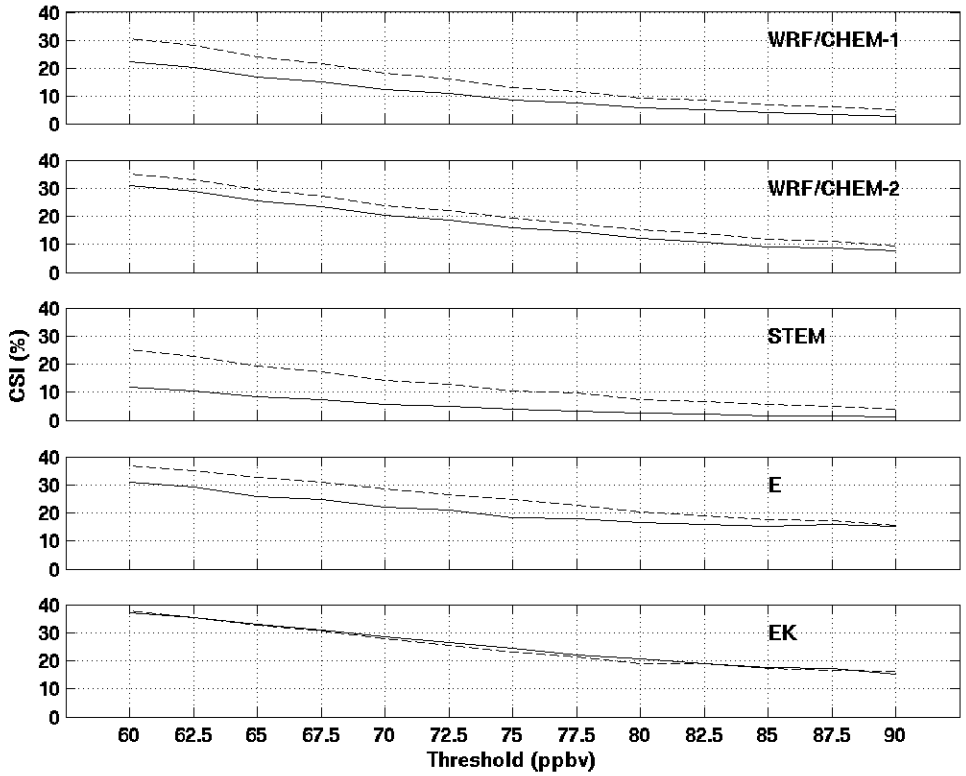

Figure 10. Similar to Figure 9, but for (from top to the bottom panel) WRF/CHEM-1, WRF/CHEM-2, STEM, the ensemble mean of the raw forecasts (E) and of the Kalman filtered forecast (EK).

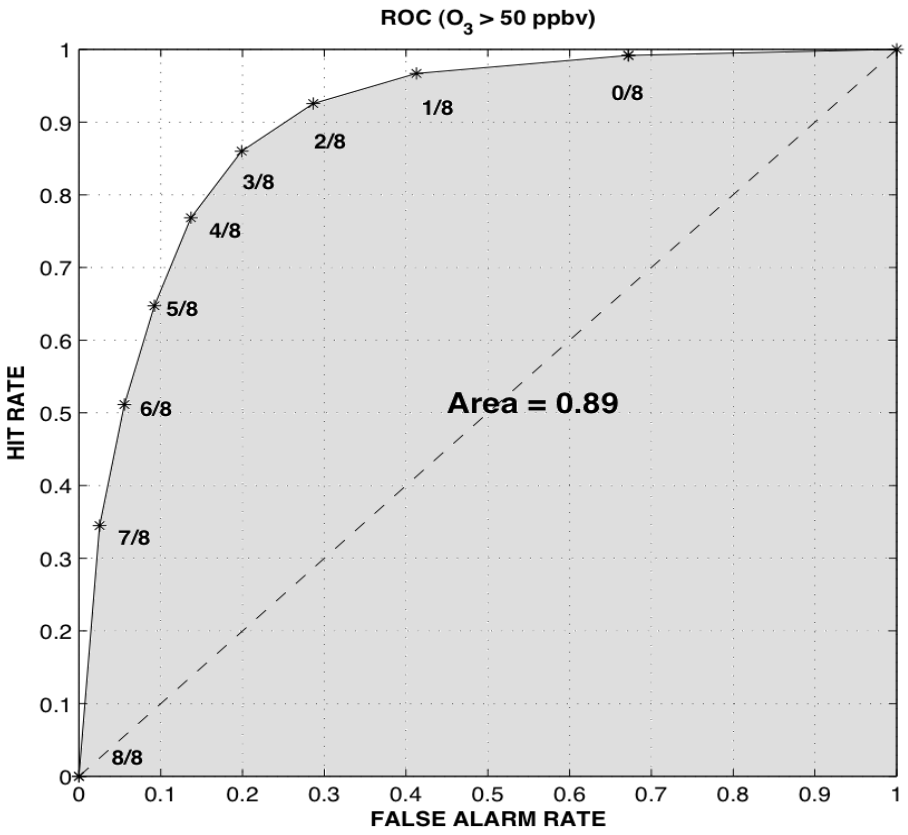

Figure 11. Relative Operating Characteristics (ROC) curve for the probabilistic forecast formed by the raw forecasts (PF-R), for observed ozone concentration above 50 ppbv. The better the probabilistic forecast, the closer the ROC curve is to the upper left corner. The shaded portion of the plot represents the ROC area (larger areas are better), and the dashed line is the ROC curve for a chance forecast. Hit rates are plotted on the ordinate against the corresponding false-alarm rates on the abscissa, to generate the ROC curve for each probability threshold (the labels adjacent to the asterisks), where the probability threshold assumes values from 0/8 to 8/8, with increments of 1/8.

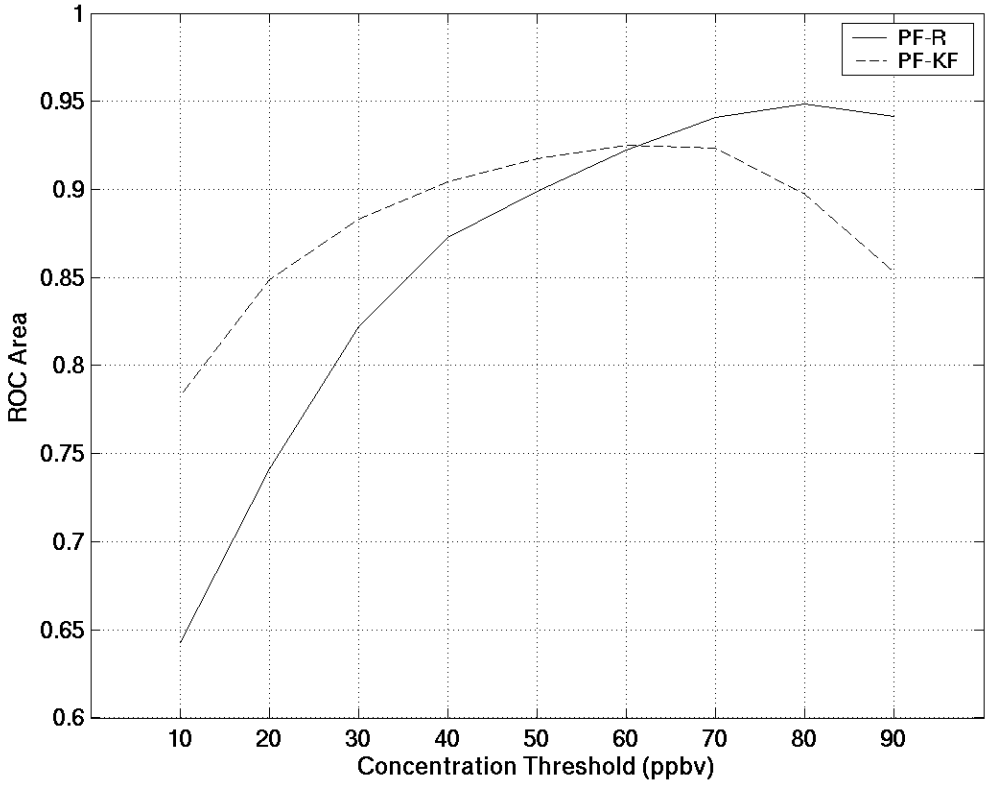

Figure 12. Relative Operating Characteristics (ROC) area values for different ozone concentration thresholds (ranging from 10 to 90 ppbv, with increments of 10 ppbv), for the probabilistic forecast formed by the raw forecasts (PF-R) (continuous line) and the probabilistic forecast formed by the Kalman-filtered corrected forecasts (PF-KF) (dashed line). Values are within the interval [0, 1], with the perfect ROC-area = 1.

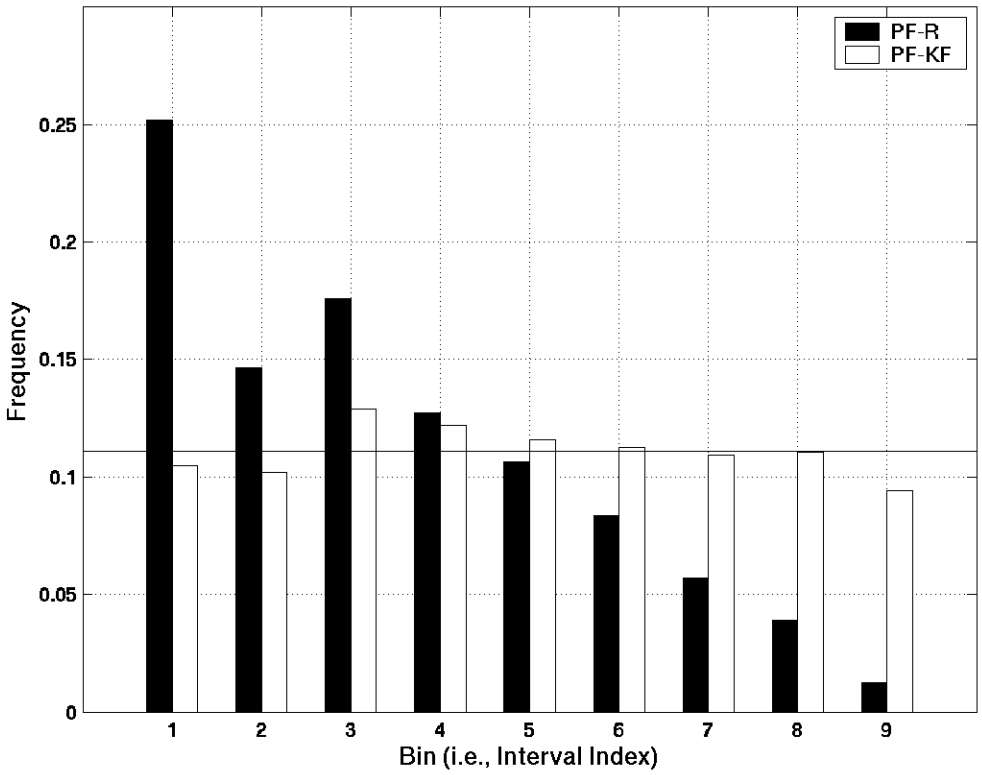

Figure 13. Talagrand diagram (rank histogram) for the probabilistic forecast formed by the raw forecasts (PF-R) (black bars) and the probabilistic forecast formed by the Kalman-filtered corrected forecasts (PF-KF) (white bars). The number of bins equals the number of ensemble members plus one. The solid horizontal line represents the perfect Talagrand diagram shape (flat). The closer the diagram to this horizontal line, the better.